Machine Learning and Text Mining Project

Submitted by : Shweta Tripathi
4/26/2024

TABLE OF CONTENT

List of figures

Part 1: Machine Learning Models

You work for an office transport company. You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding how do the employees of ABC Consulting prefer to commute presently (between home and office). Based on the parameters like age, salary, work experience etc. given in the data set 'Transport.csv', you are required to predict the preferred mode of transport. The project requires you to build several Machine Learning models and compare them so that the model can be finalized.

Data Dictionary

Age	Age of the Employee in Years
Gender	Gender of the Employee
Engineer	For Engineer =1 , Non Engineer =0
MBA	For MBA =1 , Non MBA =0
Work Exp	Experience in years
Salary	Salary in Lakhs per Annum
Distance	Distance in Kms from Home to Office
license	If Employee has Driving Licence -1, If not, then 0
Transport	Mode of Transport

1. **Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.**

Loading the dataset and checking the top 5 rows and bottom 5 rows of the dataset as follows:-

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport

Figure 1 : Top 5 rows

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

Figure 2 : Bottom 5 rows

Shape Of Data

The data contains 444 rows and 9 columns

Info Of the Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 444 entries, 0 to 443
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Age        444 non-null    int64
 1   Gender     444 non-null    object
 2   Engineer   444 non-null    int64
 3   MBA        444 non-null    int64
 4   Work Exp   444 non-null    int64
 5   Salary     444 non-null    float64
 6   Distance   444 non-null    float64
 7   license    444 non-null    int64
 8   Transport  444 non-null    object
dtypes: float64(2), int64(5), object(2)
memory usage: 31.3+ KB
```

Figure 3 : Data type of dataset

The information tells about the type of data contains in the dataset memory usages etc. The dataset contains total 9 columns among which 2 float type, 2 object type and 5 inter types.

Description of the dataset

The describe function gives the description of the dataset as follows:-

	Age	Engineer	MBA	Work Exp	Salary	Distance	license
count	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000
mean	27.747748	0.754505	0.252252	6.299550	16.238739	11.323198	0.234234
std	4.416710	0.430866	0.434795	5.112098	10.453851	3.606149	0.423997
min	18.000000	0.000000	0.000000	0.000000	6.500000	3.200000	0.000000
25%	25.000000	1.000000	0.000000	3.000000	9.800000	8.800000	0.000000
50%	27.000000	1.000000	0.000000	5.000000	13.600000	11.000000	0.000000
75%	30.000000	1.000000	1.000000	8.000000	15.725000	13.425000	0.000000
max	43.000000	1.000000	1.000000	24.000000	57.000000	23.400000	1.000000

Figure 4 : Min-Max and standard deviation

The above data tell us the min-max, standard deviation values etc. from above table we can conclude that min age is 18 whereas the max age is 43 but the mean age is approx. 28 whereas salary varies from min 6.5lpa to 57 lpa and min distance for which mode of transport is used is 3.2 kms to maximum 23.4 kms.

Duplicated values:-

The dataset does not contain any duplicate values.

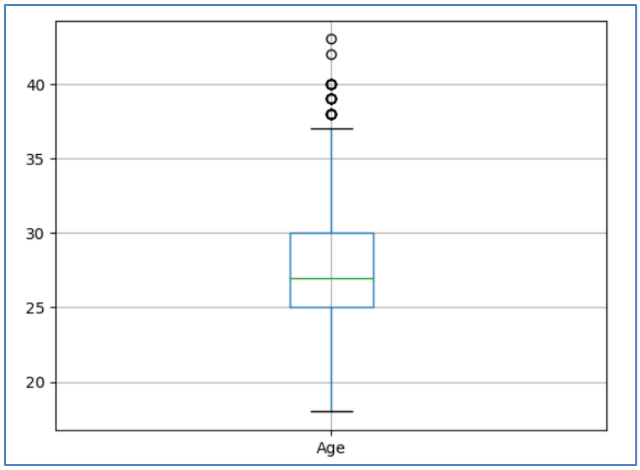
Null Values:-

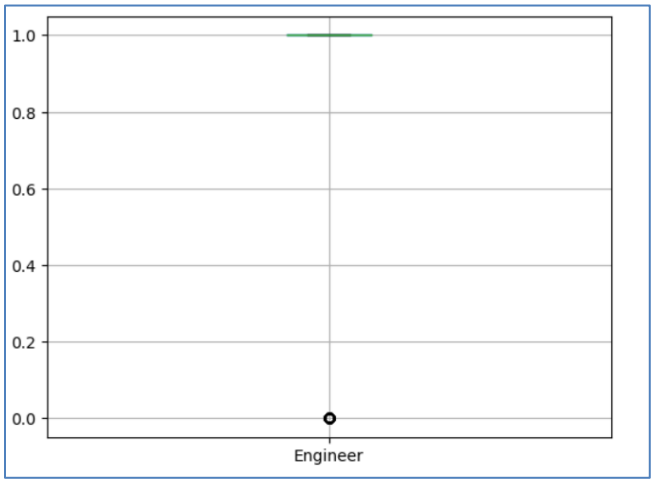
```
Age          0
Gender       0
Engineer     0
MBA          0
Work Exp     0
Salary       0
Distance     0
license      0
Transport    0
dtype: int64
```

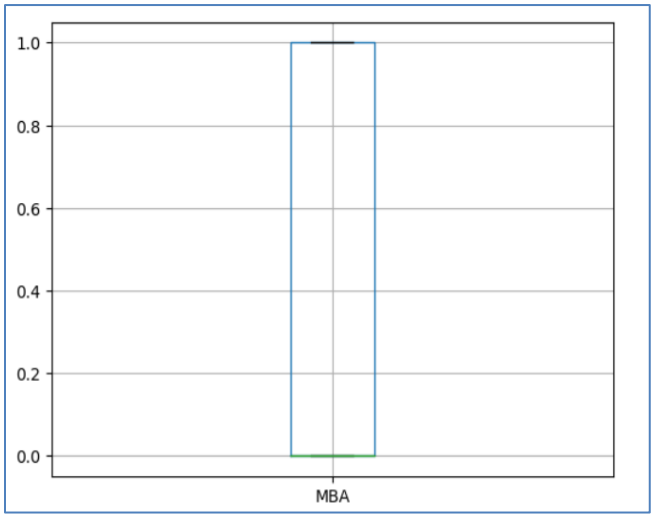
The dataset does not contain any missing values in the dataset.

Outliers Check:-

The outliers can be checked using the box as shown below


40
35
30
25
20
Age


1.0
0.8
0.6
0.4
0.2
0.0
Engineer


1.0
0.8
0.6
0.4
0.2
0.0
MBA

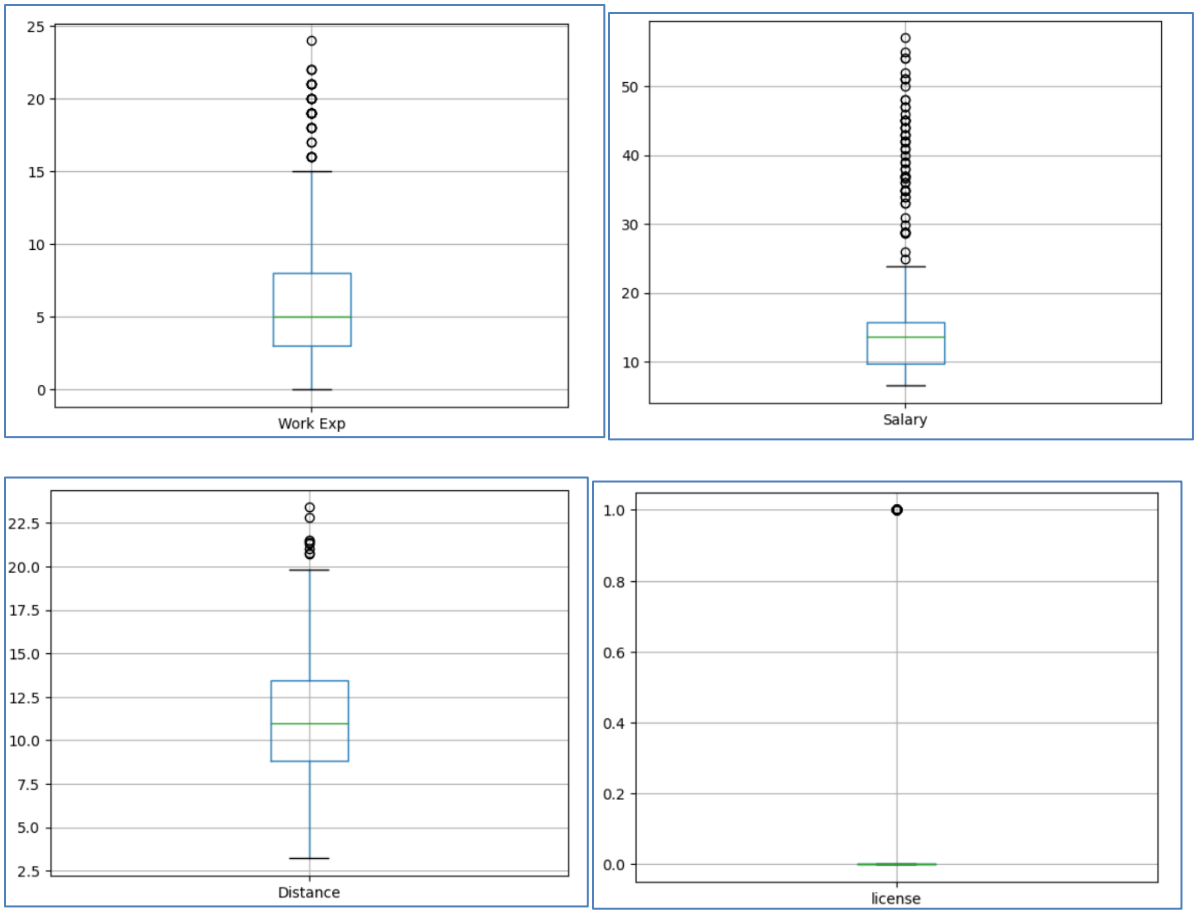
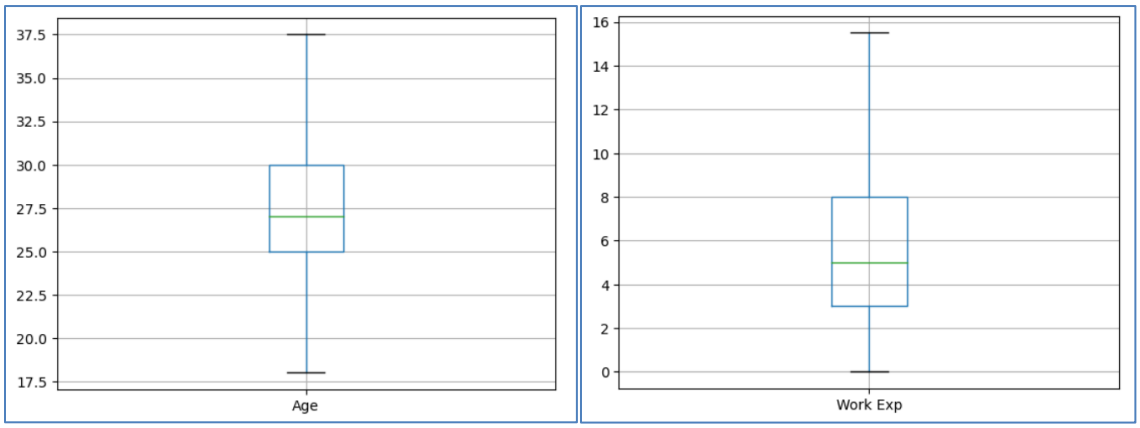


Figure 5: Box plot showing outliers

As we can observe from the above box plots that some fields contains outliers such as Age, Work Exp., Salary and Distance. Thus removing them using IQR method i.e. Mathematically, IQR = Q3 - Q1.Thus applying IQR method and removing outliers.

The box plots after removing outliers are:-



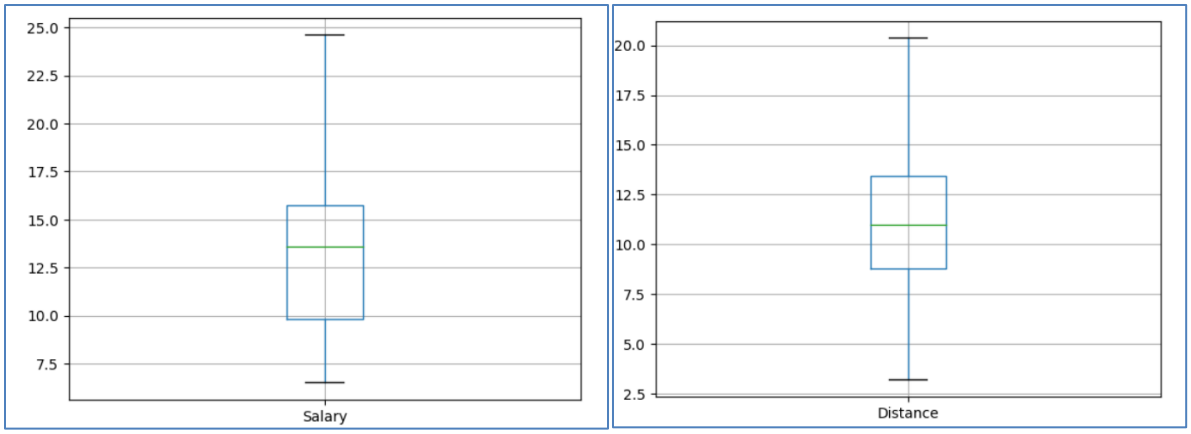


Figure 6: Box plot after removing outliers

UNIVARIATE ANALYSIS:-

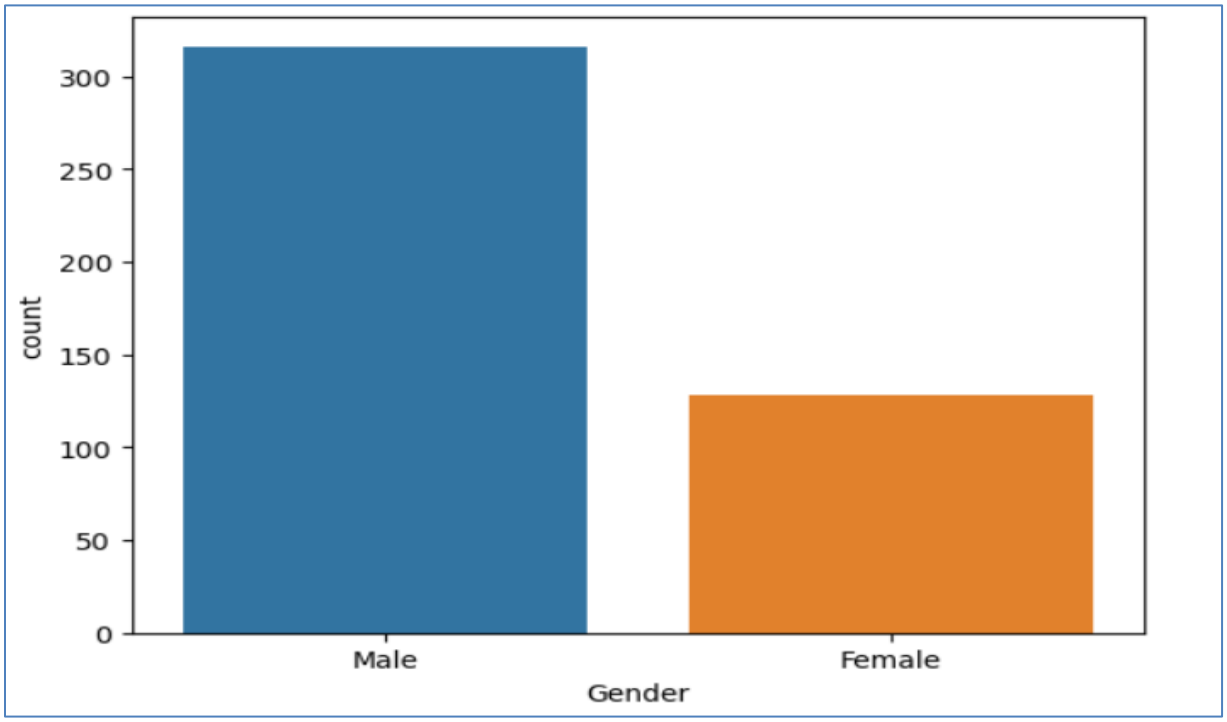
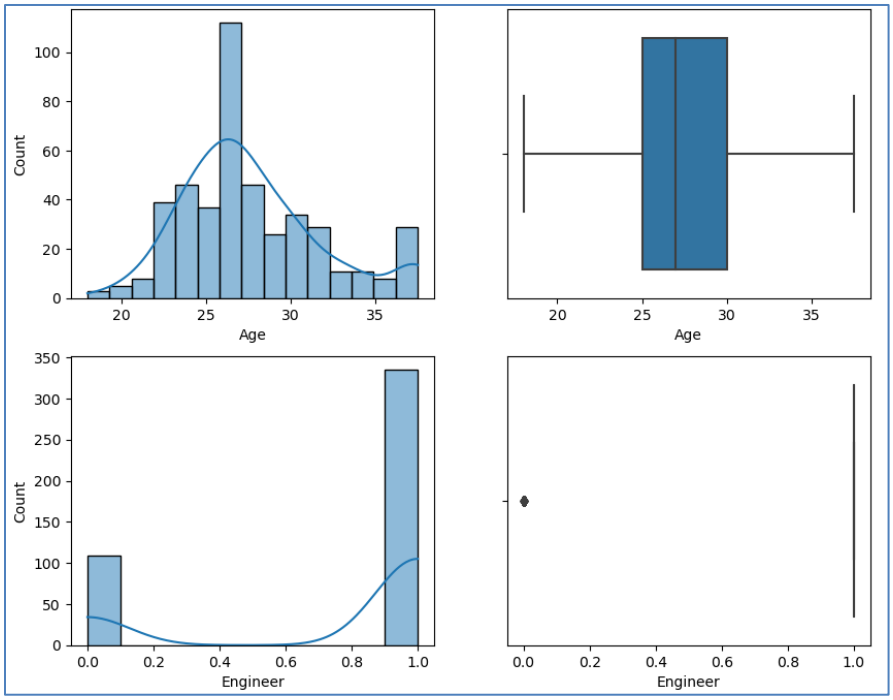
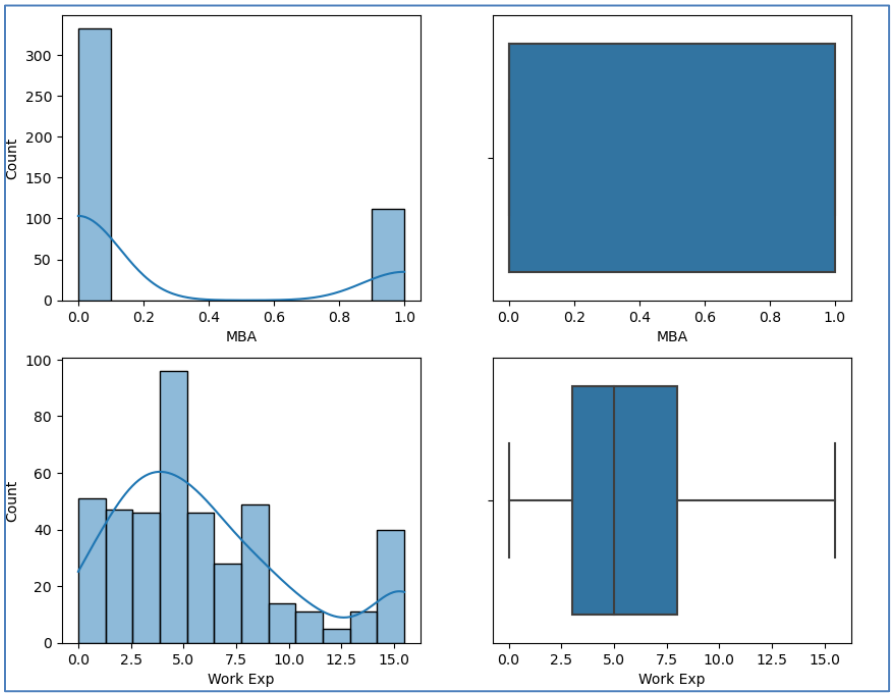


Figure 7: Univariate analysis

From the above univariate analysis of the graph it is clear that male employees are higher as compared to female employees.


Count
Age
Age
Count
Engineer
Engineer


Count
MBA
MBA
Count
Work Exp
Work Exp

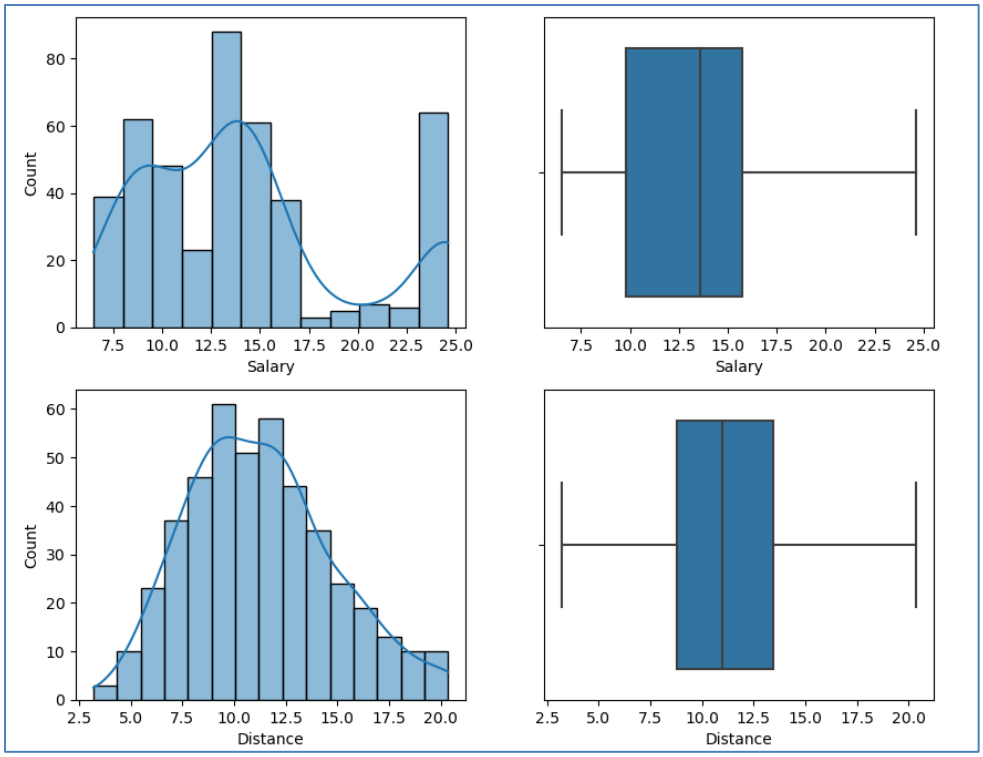
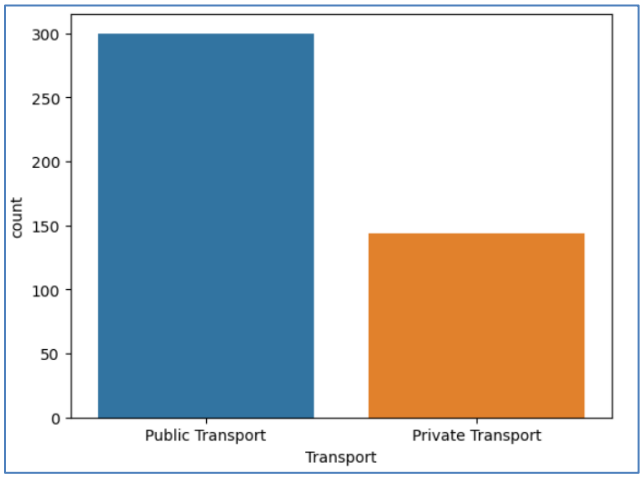


Figure 8: Univariate analysis 2



From the above univariate analysis it is clear that Public transport choose by employees higher as compared to private transport.

Bivariate Analysis:-

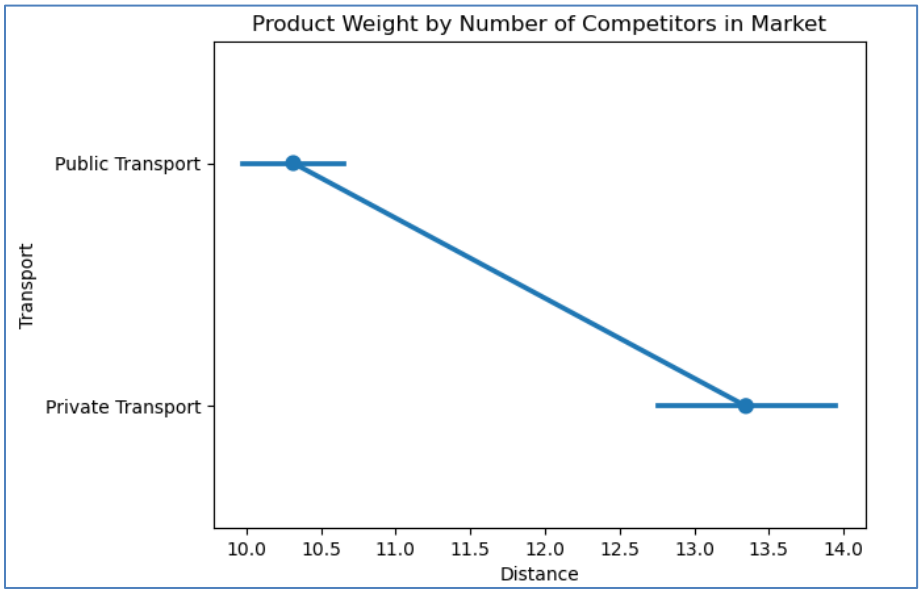


Figure 9: Bivariate analysis

The above graph shows that the below 13lpa both public and private transports are equally used but for salary above 13 lpa mostly private transport are used.

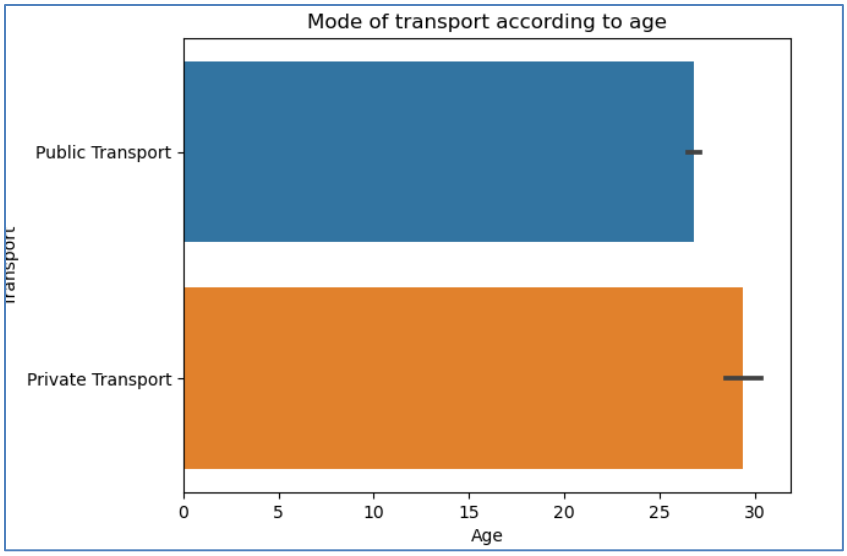


Figure 10: Bivariate analysis 2

From the above graph it is clear that higher the age more private transport used, for employees above 27 years old mostly used private transport whereas below 27 years old both transports are equally used.

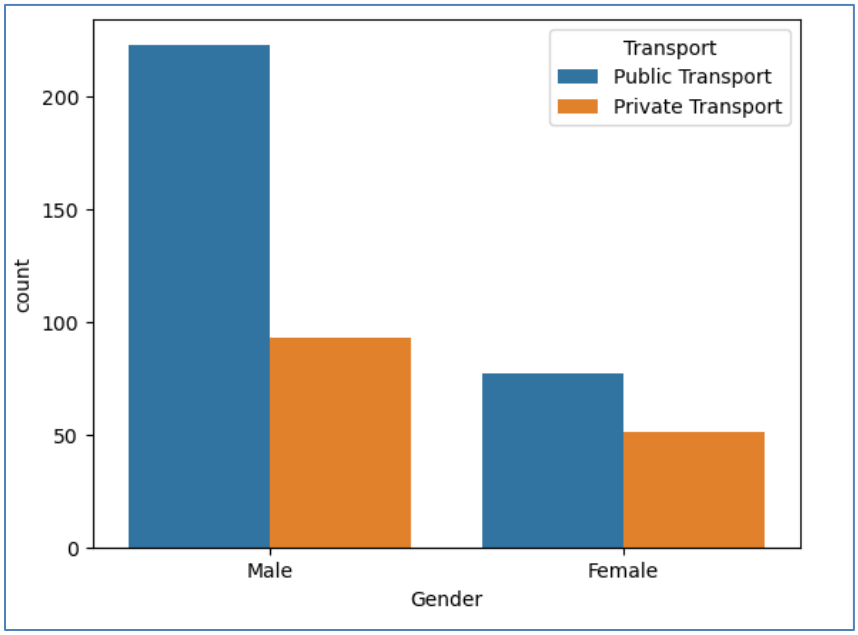


Figure 11: Bivariate Analysis 3

From the above graph it is clear that both the public transport and private transport used by male employees are higher as compared to female employees

Figure 12: Pair plot – showing positive relation

From the above pair plot we can observe that in some cases the graph is showing positive relation where as in other graphs it is not showing any relation.

Multivariate Analysis:-

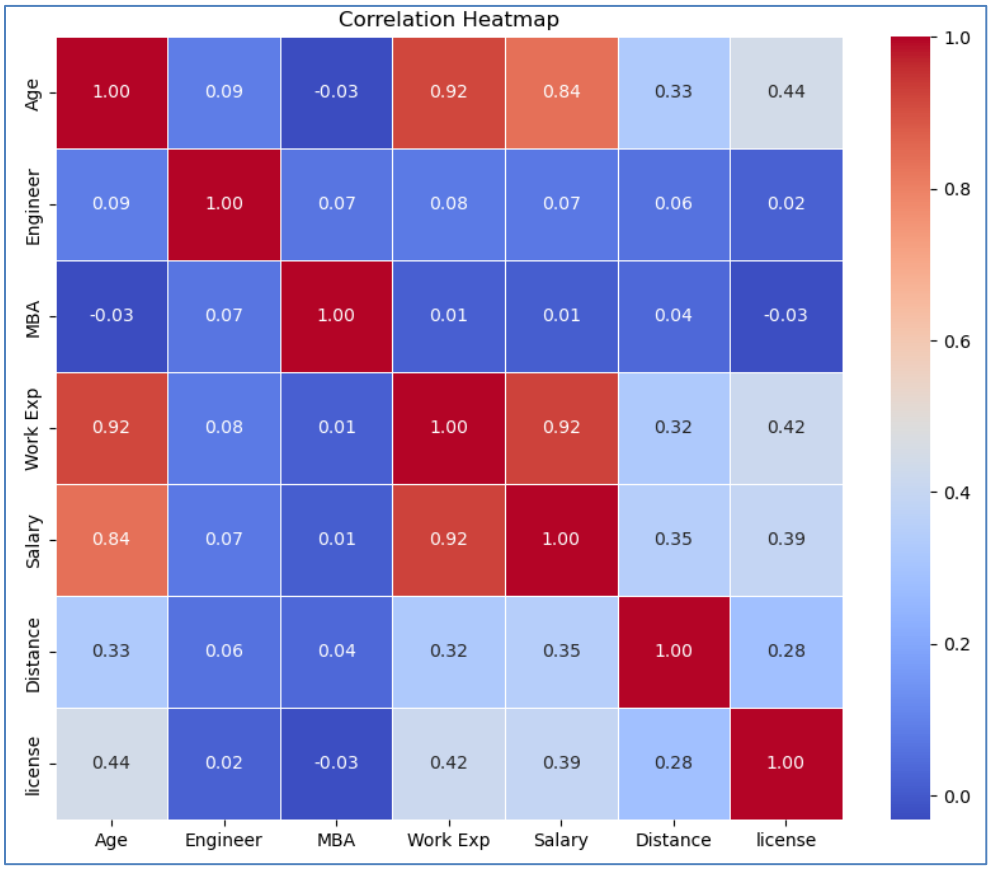


Figure 13: Multivariate analysis

From the above graph it is clear that red color shows high correlation whereas dark blue color shows least correlation like work experience, salary, age shows high correlation to each other whereas license and engineer as very less correlated.

Scaling the data:-

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	0.083578	Male	-1.753110	-0.580818	-0.459062	0.030061	-2.292795	-0.553066	Public Transport
1	-1.117345	Female	0.570415	-0.580818	-0.459062	-1.098453	-2.264482	-0.553066	Public Transport
2	0.323762	Male	0.570415	-0.580818	0.235791	-0.139216	-2.037981	-0.553066	Public Transport
3	0.083578	Female	0.570415	1.721710	-0.227444	-0.139216	-1.924730	-0.553066	Public Transport
4	-0.156607	Male	0.570415	-0.580818	-0.459062	-0.139216	-1.896417	-0.553066	Public Transport

Figure 14: Scaling data

The StandardScaler scaling method is used to provided dataset consists of scaled attributes such as age, gender, employment-related factors, salary, distance, and transportation mode, likely prepared for machine learning analysis, with values normalized or scaled for uniformity across features.

2. Split the data into train and test in the ratio 70:30. Is scaling necessary or not?

Train Test Split:-

This method separates predictor variables (features) from the target variable. Predictor variables are stored in the dataframe('X') by dropping the 'Transport' column from the original dataframe, while the target variable is stored in the dataframe ('y').

We imported train_test_split function from scikit-learn to split the data into training and testing sets. The data is split into 70% training and 30% testing sets, ensuring that the class distribution of the target variable ('Transport') is maintained in both sets. The random state is set to 1 for reproducibility.

Yes Scaling is necessary in machine learning to ensure that features have similar scales, improving algorithm performance and convergence.

3. Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.

a. Logistic Regression Model

b. Linear Discriminant Analysis

c. Decision Tree Classifier – CART model

d. Naïve Bayes Model

e. KNN Model

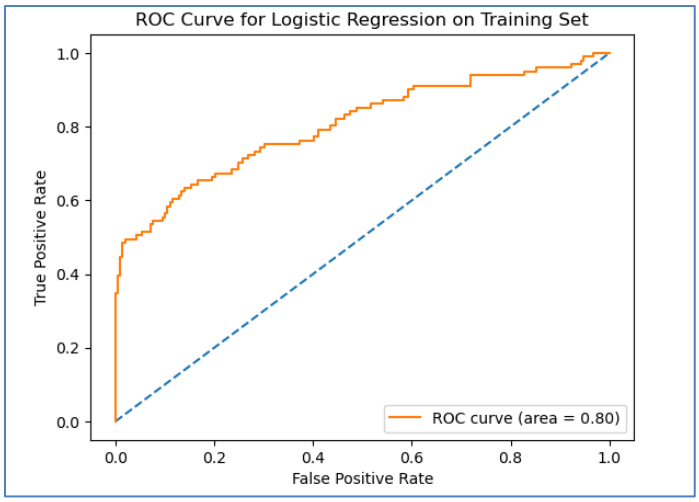
f. Random Forest Model

g. Boosting Classifier Model using Gradient boost.

Logistic Regression Model

Training model evaluation:-

```
0.7903225806451613
```



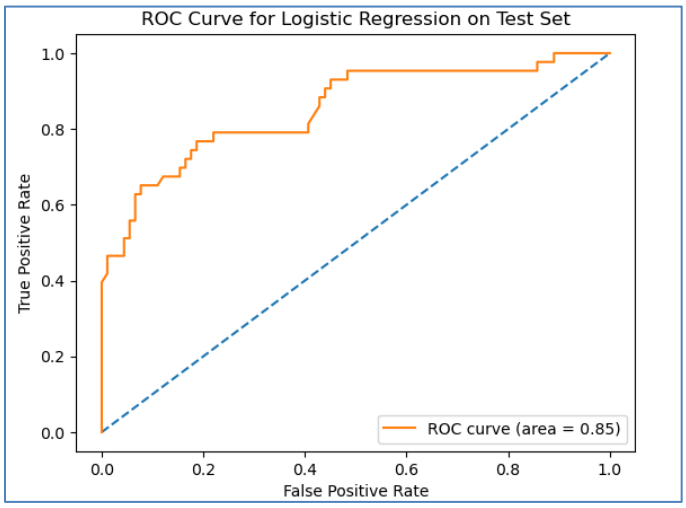


Figure 15: ROC Curve – Training and Test Set - LR

An AUC of 0.80 for training and 0.85 for testing suggests that the logistic regression model performs well in both training and testing datasets. However, the slightly higher AUC in the testing dataset indicates that the model generalizes well to unseen data, which is a positive sign of its robustness. This suggests that the model is likely not over fitting and is effectively capturing the underlying patterns in the data.

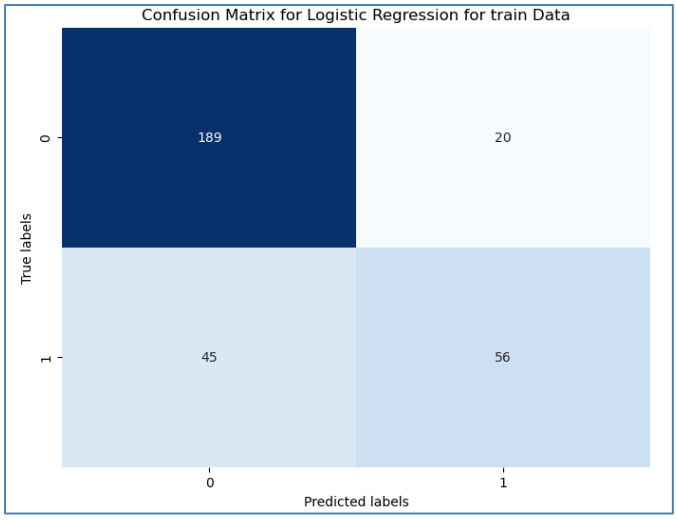


Figure 16: Confusion Matrix Train Data - LR

With 189 true negatives and 56 true positives, the model correctly predicted 245 instances out of 310. It misclassified 65 instances, with 45 false negatives and 20 false positives.

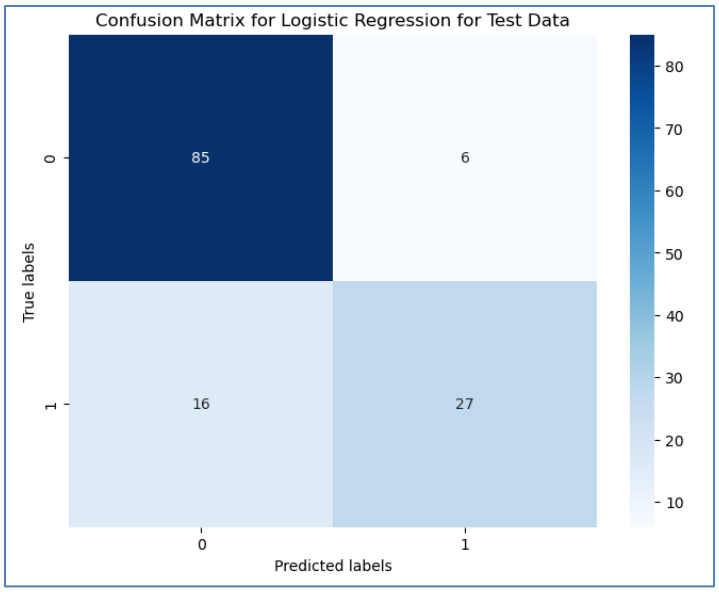


Figure 17: Confusion Matrix Test data - LR

The model correctly predicted 112 instances out of 134, with 85 true negatives and 27 true positives. It misclassified 22 instances, with 16 false negatives and 6 false positives.

The model appears to generalize well to unseen data, as indicated by its performance on the testing dataset being comparable to or slightly better than its performance on the training dataset.

	precision	recall	f1-score	support
0	0.81	0.90	0.85	209
1	0.74	0.55	0.63	101
accuracy			0.79	310
macro avg	0.77	0.73	0.74	310
weighted avg	0.78	0.79	0.78	310

	precision	recall	f1-score	support
0	0.84	0.93	0.89	91
1	0.82	0.63	0.71	43
accuracy			0.84	134
macro avg	0.83	0.78	0.80	134
weighted avg	0.83	0.84	0.83	134

The classification tables provide detailed performance metrics for both training and testing datasets, evaluating the logistic regression model's ability to correctly classify instances into their respective classes (0 and 1)

LDA:-

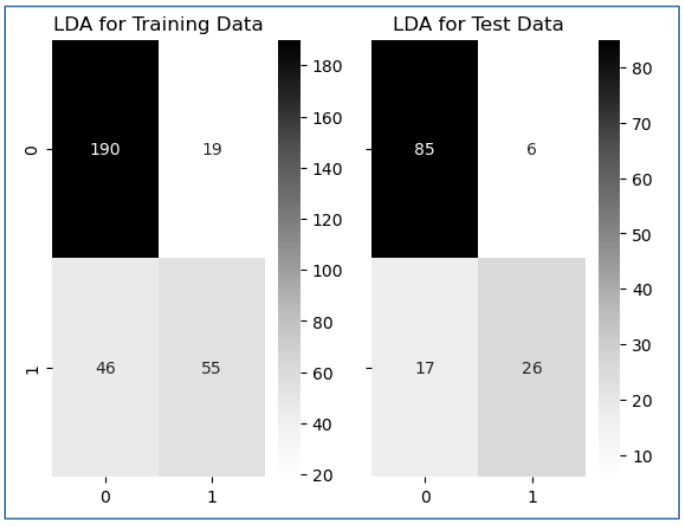


Figure 18: LDA – Training and Test Data

The confusion matrices for Linear Discriminant Analysis (LDA) show the counts of correct and incorrect predictions made by the model on both training and testing datasets. They help

evaluate the model's performance in distinguishing between classes, with higher counts along the diagonal (true positives and true negatives) indicating better classification accuracy.

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.81	0.91	0.85	209
1	0.74	0.54	0.63	101
accuracy			0.79	310
macro avg	0.77	0.73	0.74	310
weighted avg	0.78	0.79	0.78	310

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.83	0.93	0.88	91
1	0.81	0.60	0.69	43
accuracy			0.83	134
macro avg	0.82	0.77	0.79	134
weighted avg	0.83	0.83	0.82	134

The classification reports for Linear Discriminant Analysis (LDA) on both training and testing data show performance metrics such as precision, recall, and F1-score for each class (0 and 1), as well as overall accuracy. LDA demonstrates relatively good accuracy and precision

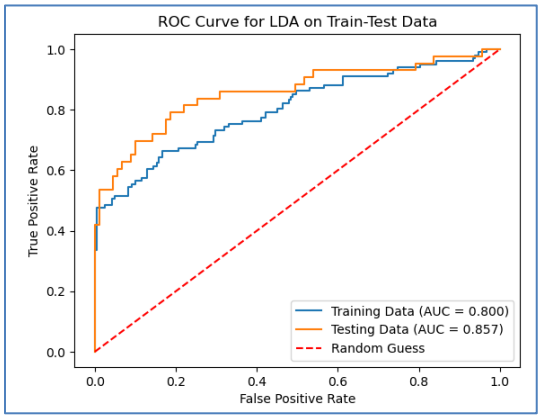


Figure 19: ROC Curve – LDA Analysis

AUC curve for both training and testing data suggest that the model performs well in distinguishing between classes, with an AUC of 0.800 for training and 0.857 for testing. These values indicate that LDA effectively separates the classes, demonstrating good discrimination

ability. The higher AUC value on the testing data suggests that the model generalizes well to unseen data, further supporting its effectiveness in classification tasks.

Decision Tree Classifier

```
DecisionTreeClassifier
DecisionTreeClassifier(ccp_alpha=0.001, criterion='entropy', max_depth=5,
                       max_features='sqrt', min_samples_leaf=5,
                       random_state=1024)
```

The Decision Tree Classifier is configured to use entropy as the criterion for splitting nodes, aiming to maximize information gain. With a maximum tree depth of 5 levels, it prevents over fitting by limiting the complexity of the model. Additionally, it employs square root of the total number of features for splitting at each node and sets a minimum of 5 samples per leaf, ensuring leaf nodes contain enough instances for robust decision-making.

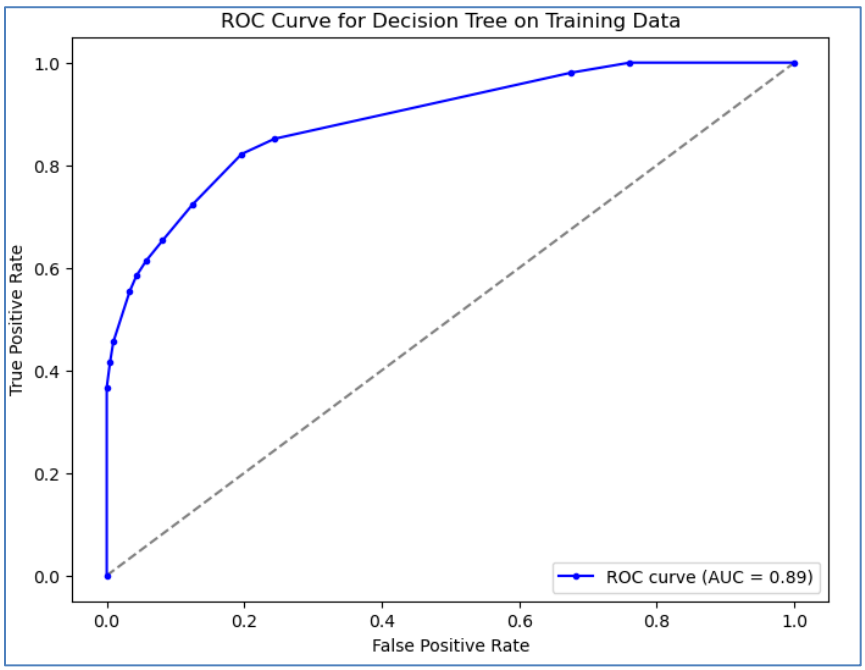


Figure 20: ROC Curve Decision Tree

	precision	recall	f1-score	support
0	0.83	0.96	0.89	209
1	0.87	0.58	0.70	101
accuracy			0.84	310
macro avg	0.85	0.77	0.79	310
weighted avg	0.84	0.84	0.83	310

The ROC curve's AUC (Area under the Curve) value of 0.89 for the Decision Tree model suggests strong performance in distinguishing between classes. With higher AUC values indicating better discrimination ability.

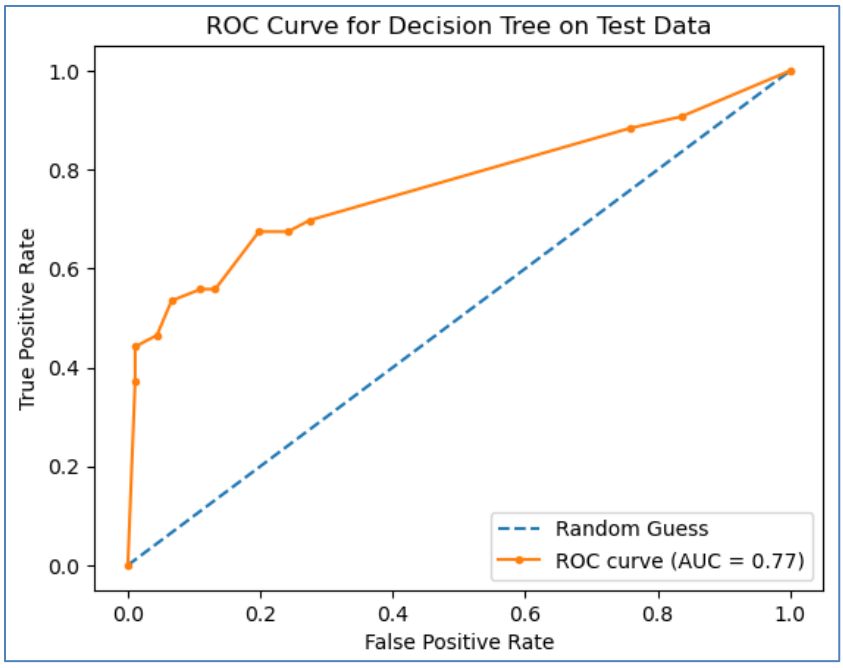


Figure 21 : ROC Curve – Test data – Decision Tree

	precision	recall	f1-score	support
0	0.81	0.93	0.87	91
1	0.79	0.53	0.64	43
accuracy			0.81	134
macro avg	0.80	0.73	0.75	134
weighted avg	0.80	0.81	0.79	134

A ROC curve with an AUC (Area under the Curve) value of 0.77 for the Decision Tree model on testing data suggests reasonable but not exceptional performance in distinguishing between classes. While an AUC of 0.77 indicates that the model has some ability to discriminate between classes, it may not perform as well as desired compared to other models.

Gaussian Naive Bayes

```
0.7935483870967742
[[191  18]
 [ 46  55]]
              precision    recall  f1-score   support

           0       0.81      0.91      0.86       209
           1       0.75      0.54      0.63       101

    accuracy                           0.79       310
   macro avg       0.78      0.73      0.74       310
weighted avg       0.79      0.79      0.78       310
```

Figure 22: Confusion matrix – Gaussian naïve Bayes

The classification for the Gaussian Naive Bayes model on the training data shows accuracy of around 79.35%. The confusion matrix indicates that the model correctly classified a majority of instances, particularly for class 0, with 191 true positives and 18 false positives.

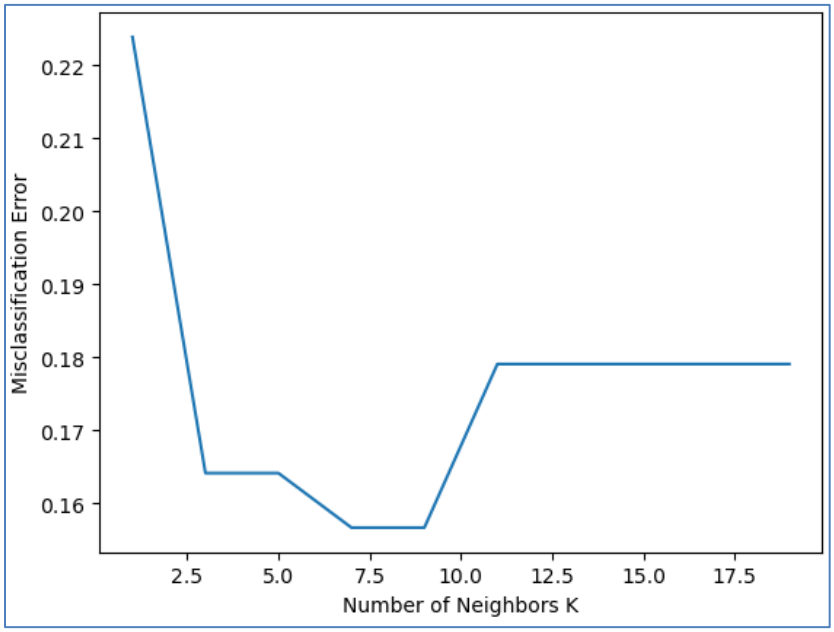
```
0.7985074626865671
[[83  8]
 [19 24]]
              precision    recall  f1-score   support

           0       0.81      0.91      0.86        91
           1       0.75      0.56      0.64        43

    accuracy                           0.80       134
   macro avg       0.78      0.74      0.75       134
weighted avg       0.79      0.80      0.79       134
```

The provided metrics for the Gaussian Naive Bayes model on the testing data reveal an accuracy of approximately 79.85%. The confusion matrix illustrates that out of 134 instances, 83 were correctly classified as class 0, and 24 were correctly classified as class 1, with 8 instances falsely classified as class 1 and 19 falsely classified as class 0.

KNN Model

```
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=11)
```

K-Nearest Neighbors (KNN) classifier with k=11 neighbors indicates that the model predicts the class of a data point by considering the majority class among its 11 nearest neighbors in the feature space.




```
0.8129032258064516
[[201   8]
 [ 50  51]]
              precision    recall  f1-score   support

           0       0.80      0.96      0.87       209
           1       0.86      0.50      0.64       101

    accuracy                           0.81       310
   macro avg       0.83      0.73      0.76       310
weighted avg       0.82      0.81      0.80       310
```

Figure 23– Confusion matrix – KNN model

The K-Nearest Neighbors (KNN) model on the training data gives an accuracy of approximately 81.29%. The confusion matrix shows that out of 310 instances, 201 were correctly classified as class 0, and 51 were correctly classified as class 1.

```
0.8208955223880597
[[88  3]
 [21 22]]
              precision    recall  f1-score   support

           0       0.81      0.97      0.88        91
           1       0.88      0.51      0.65        43

    accuracy                           0.82       134
   macro avg       0.84      0.74      0.76       134
weighted avg       0.83      0.82      0.81       134
```

The K-Nearest Neighbors (KNN) model on the testing data indicates an accuracy of approximately 82.09%. The confusion matrix shows that out of 134 instances, 88 were correctly classified as class 0, and 22 were correctly classified as class 1.

Random Forest

```
RandomForestClassifier
RandomForestClassifier(random_state=1)
```

Train model prediction score

```
1.0
[[209   0]
 [  0 101]]
```

The provided metrics for the Random Forest model on the training data indicate a perfect accuracy of 100%. The confusion matrix shows that out of 310 instances, all 209 instances of class 0 and all 101 instances of class 1 were correctly classified. The perfect performance on the training data may indicate potential over fitting.

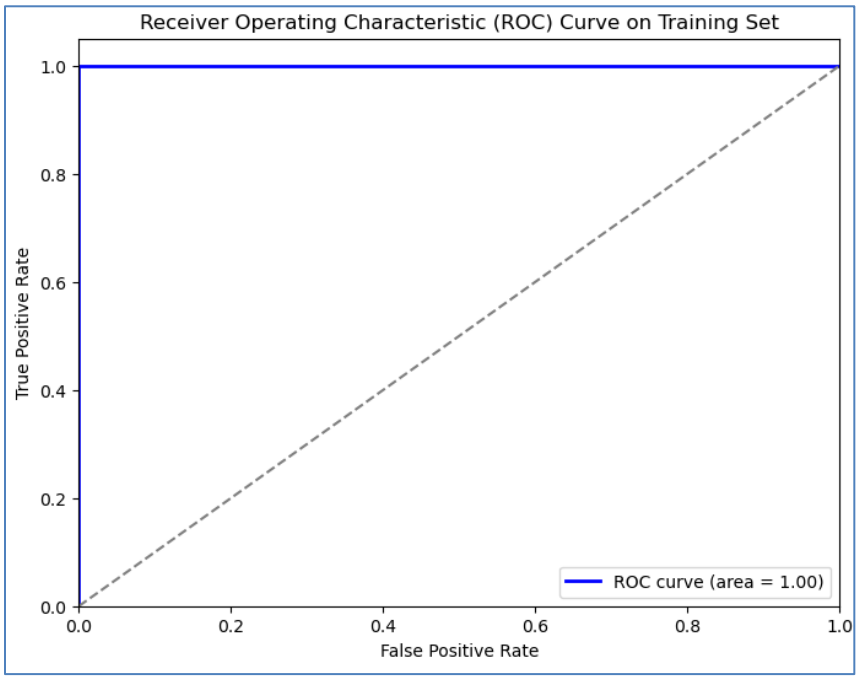


Figure 24: ROC Curve – Random Forest

```
0.8731343283582089
[[87  4]
 [13 30]]
              precision    recall  f1-score   support

           0       0.87      0.96      0.91        91
           1       0.88      0.70      0.78        43

    accuracy                           0.87       134
   macro avg       0.88      0.83      0.85       134
weighted avg       0.87      0.87      0.87       134
```

The provided metrics for the Random Forest model on the testing data indicate an accuracy of approximately 87.31%. The confusion matrix shows that out of 134 instances, 87 were correctly classified as class 0, and 30 were correctly classified as class 1. The model demonstrates relatively good performance, with high precision, recall, and F1-score for both classes.

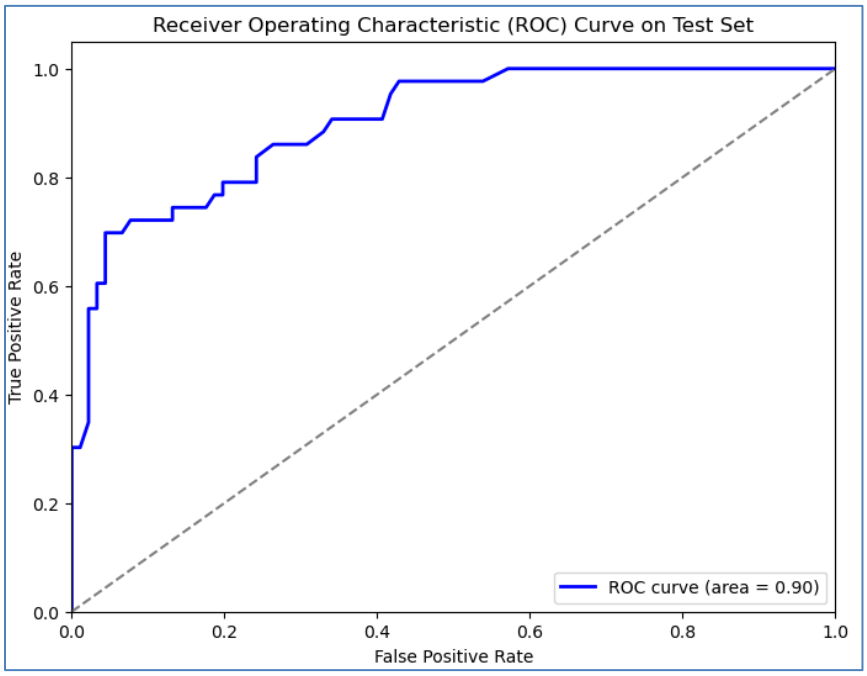


Figure 25: ROC Curve – Random Forest – Test set

Over-fitting

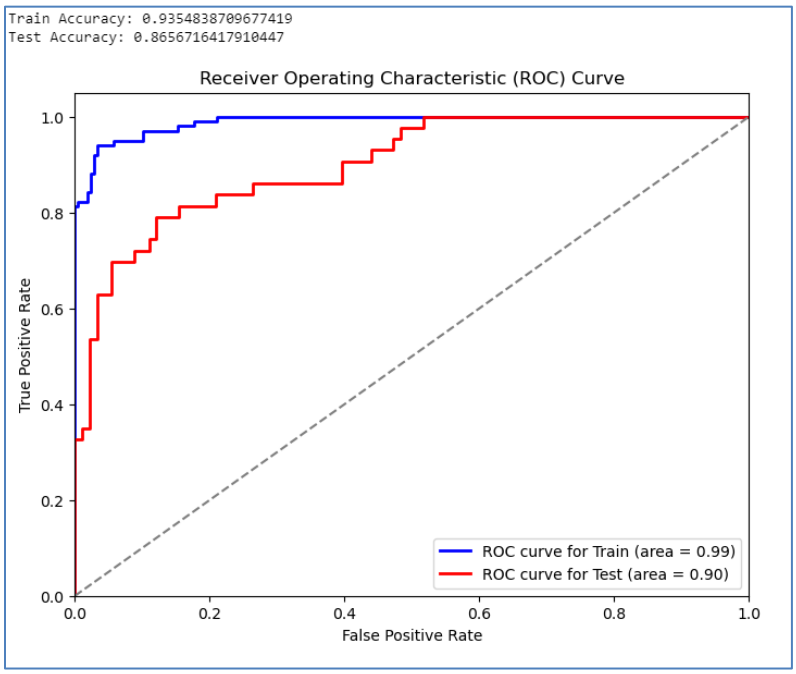


Figure 26: ROC Curve - over fitting

The model achieves a high accuracy of approximately 93.55% on the training data and 86.57% on the testing data. This indicates that the model effectively classifies the majority of instances correctly on both datasets.

The ROC curve AUC (Area under the Curve) values are high, with 0.99 for the training data and 0.90 for the testing data. These values indicate excellent discrimination ability, suggesting that the model can effectively distinguish between the two classes.

Ensemble Learning – Gradient Boost

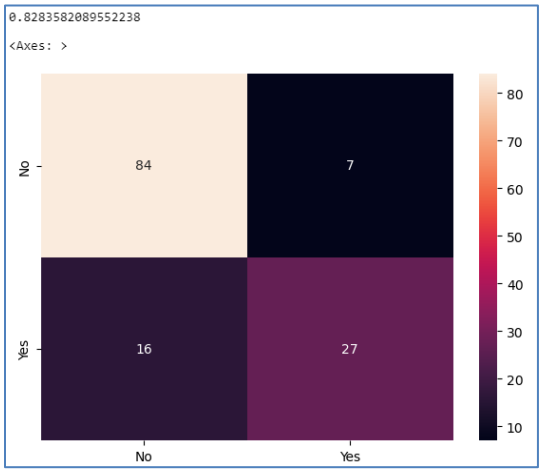


Figure 27: Confusion matrix – Ensemble learning

The confusion matrix shows that out of 134 instances, 84 were correctly classified as class 0, and 27 were correctly classified as class 1. Additionally, 16 instances of class 0 were incorrectly classified as class 1, and 7 instances of class 1 were incorrectly classified as class 0.

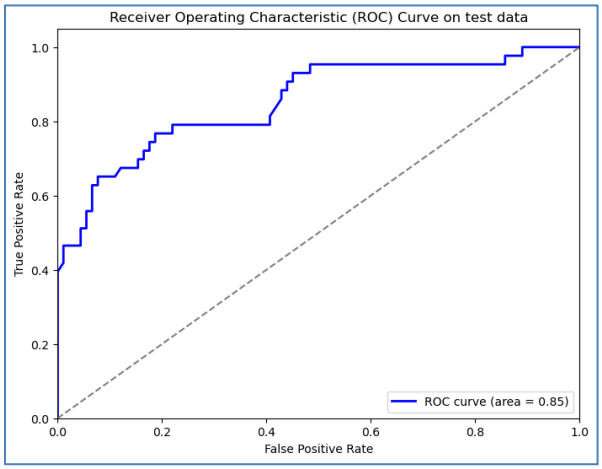


Figure 28: ROC Curve – Test data – Ensemble learning

The ROC curve AUC (Area under the Curve) value is 0.85, indicating good discrimination ability of the model in distinguishing between the classes.

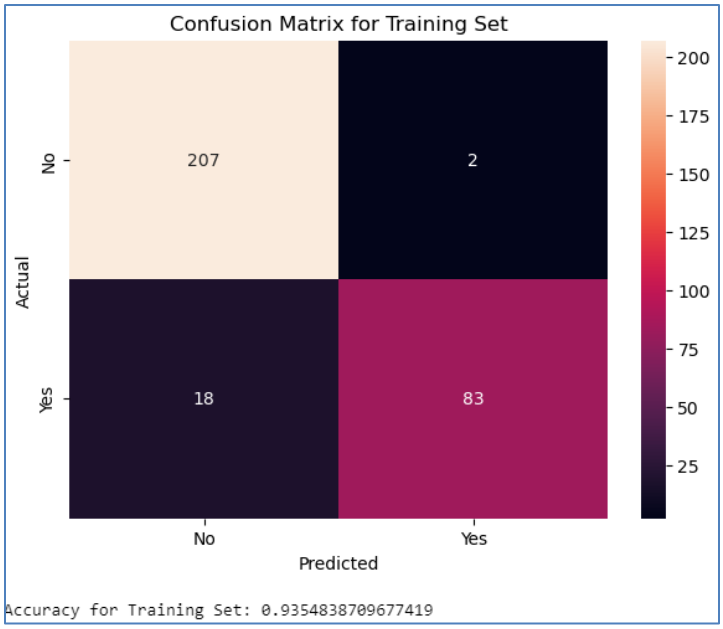


Figure 29: confusion matrix – training set – ensemble learning

The confusion matrix reveals that out of 310 instances, 207 were correctly classified as class 0, and 83 were correctly classified as class 1. Additionally, 18 instances of class 0 were incorrectly classified as class 1, and 2 instances of class 1 were incorrectly classified as class 0. This demonstrates the model's ability to effectively distinguish between the two classes, with a high number of correct classifications.

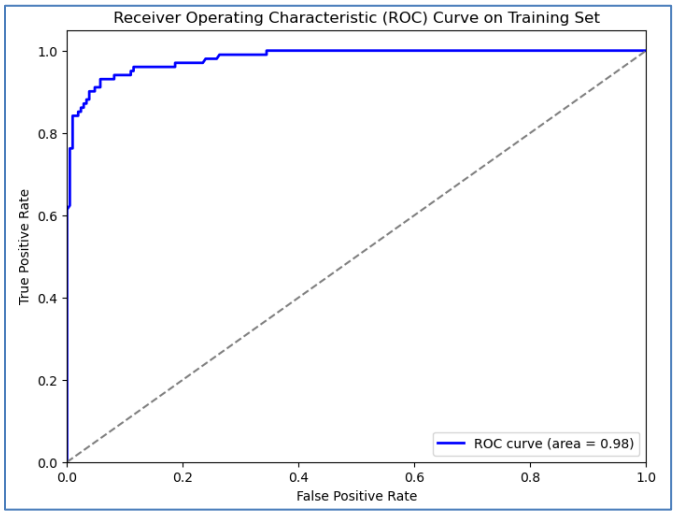


Figure 30: ROC Curve – Training set – Ensemble learning

The ROC curve AUC (Area under the Curve) value is 0.98, indicating exceptional discrimination ability of the model. This suggests that the Gradient Boosting model can distinguish between the classes with high accuracy.

4. Which model performs the best?

Random Forest achieves a high accuracy of 93.55% on the training data and 86.57% on the testing data, with ROC curve AUC values of 0.99 and 0.90, respectively.

The Random Forest thus appears to perform the best due to their higher accuracies and strong discrimination abilities indicated by the ROC curve AUC values.

5. What are your business insights?

- Gender plays a significant role in transport preferences, with male employees generally outnumbering female employees. Additionally, public transport is preferred over private transport by the majority of employees. These insights can inform transport service providers about the target demographic and their preferences, helping tailor transportation services accordingly
- There is a noticeable shift towards private transport for employees earning above a certain salary threshold, indicating that higher income levels enable employees to afford and prefer private transportation options. Employers can consider offering transportation benefits or incentives tailored to employees' salary levels to accommodate their preferences
- Older employees, particularly those above 27 years old, tend to use private transport more frequently compared to younger employees. This suggests that age influences transportation preferences, with older employees potentially valuing factors such as comfort, convenience, and flexibility offered by private transport options.
- The analysis of machine learning models provides valuable insights into their performance in predicting employees' transport preferences. Models with high accuracy, precision, recall, and AUC-ROC values, such as Random Forest and Gradient Boosting, can help transport service providers make informed decisions about resource allocation, route planning, and service optimization to meet the needs and preferences of their target audience effectively.
- By understanding the factors influencing employees' transport preferences, businesses can segment their customer base and tailor marketing strategies to target specific demographic groups more effectively. For example, offering promotions or discounts on private transport services to high-income earners or older employees may attract more customers and increase revenue.
- Transport service providers should continuously monitor and evaluate customer feedback, market trends, and competitor offerings to adapt their services and stay competitive. Regular updates and refinements to transportation infrastructure, scheduling, and amenities can enhance customer satisfaction and loyalty, driving long-term business success.

Part 2: Text Mining

A dataset of Shark Tank episodes is made available. It contains 495 entrepreneurs making their pitch to the VC sharks.

You will ONLY use “Description” column for the initial text mining exercise.

The Top 5 rows of the dataset are as following:-

	deal	description	episode	category	entrepreneurs	location	website	askedFor	exchangeForStake	valuation	season	shark1
0	False	Bluetooth device implant for your ear.	1	Novelties	Darrin Johnson	St. Paul, MN	NaN	1000000	15	6666667	1	Barbara Corcoran
1	True	Retail and wholesale pie factory with two reta...	1	Specialty Food	Tod Wilson	Somerset, NJ	http://whybake.com/	460000	10	4600000	1	Barbara Corcoran
2	True	Ava the Elephant is a godsend for frazzled par...	1	Baby and Child Care	Tiffany Krumins	Atlanta, GA	http://www.avatheelephant.com/	50000	15	333333	1	Barbara Corcoran
3	False	Organizing, packing, and moving services deliv...	1	Consumer Services	Nick Friedman, Omar Soliman	Tampa, FL	http://collegehunkshaulingjunk.com/	250000	25	1000000	1	Barbara Corcoran
4	False	Interactive media centers for healthcare waiti...	1	Consumer Services	Kevin Flannery	Cary, NC	http://www.wispots.com/	1200000	10	12000000	1	Barbara Corcoran

Figure 31: Top 5 rows – Text mining

Shape:-

Total rows in the dataset are 495 whereas total columns are 19

Null values:-

The null values in the dataset can be checked by isnull function which give the following result:-

```
deal                       0
description                0
episode                    0
category                   0
entrepreneurs             72
location                   0
website                   38
askedFor                   0
exchangeForStake           0
valuation                  0
season                     0
shark1                     0
shark2                     0
shark3                     0
shark4                     0
shark5                     0
title                      0
episode-season             0
Multiple Entreprenuers     0
dtype: int64
```

From the above columns it is clear that entrepreneurs and website contains null values thus removing the null values than we get the following result

```
deal                       0
description                0
episode                    0
category                   0
entrepreneurs              0
location                   0
website                    0
askedFor                   0
exchangeForStake           0
valuation                  0
season                     0
shark1                     0
shark2                     0
shark3                     0
shark4                     0
shark5                     0
title                      0
episode-season             0
Multiple Entreprenuers     0
dtype: int64
```

1. **Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.**

```
      deal                                        description
1     True  Retail and wholesale pie factory with two reta...
2     True  Ava the Elephant is a godsend for frazzled par...
3    False  Organizing, packing, and moving services deliv...
4    False  Interactive media centers for healthcare waiti...
5     True  One of the first entrepreneurs to pitch on Sha...
..     ...                                               ...
490   True  Zoom Interiors is a virtual service for interi...
491   True  Spikeball started out as a casual outdoors gam...
492   True  Shark Wheel is out to literally reinvent the w...
493  False  Adriana Montano wants to open the first Cat Ca...
494   True  Sway Motorsports makes a three-wheeled, all-el...

[387 rows x 2 columns]
```

2. **Create two corpora, one with those who secured a Deal, the other with those who did not secure a deal.**

<u>Secured Data :-</u>

	deal	description	episode	category	entrepreneurs	location	website	askedFor	exchangeForStake	valuation	season	shark1	
1	True	Retail and wholesale pie factory with two reta...	1.0	Specialty Food	Tod Wilson	Somerset, NJ	http://whybake.com/	460000.0	10.0	4600000.0	1.0	Barbara Corcoran	H
2	True	Ava the Elephant is a godsend for frazzled par...	1.0	Baby and Child Care	Tiffany Krumins	Atlanta, GA	http://www.avatheelephant.com/	50000.0	15.0	333333.0	1.0	Barbara Corcoran	H
5	True	One of the first entrepreneurs to pitch on Sha...	2.0	Specialty Food	Susan Knapp	Napa Valley, CA	http://www.aperfectpear.com	500000.0	15.0	3333333.0	1.0	Barbara Corcoran	H
12	True	A line of books written to help children find ...	3.0	Baby and Child Care	Lori Lite	Marietta, GA	http://www.stressfreekids.com	250000.0	20.0	1250000.0	1.0	Barbara Corcoran	H
16	True	Coverplay is a slipcover for children's play y...	4.0	Baby and Child Care	Amy Feldman and Allison Costa	Los Angeles, CA	http://www.coverplayard.com/	350000.0	15.0	2333333.0	1.0	Barbara Corcoran	H

Figure 32: Secured data

Not secured Data:-

	deal	description	episode	category	entrepreneurs	location	website	askedFor	exchangeForStake	valuation	season	shar
3	False	Organizing, packing, and moving services deliv...	1.0	Consumer Services	Nick Friedman, Omar Soliman	Tampa, FL	http://collegehunkshaulingjunk.com/	250000.0	25.0	1000000.0	1.0	Barb Corco
4	False	Interactive media centers for healthcare waiti...	1.0	Consumer Services	Kevin Flannery	Cary, NC	http://www.wispots.com/	1200000.0	10.0	12000000.0	1.0	Barb Corco
6	False	A mixed martial arts clothing line looking to ...	2.0	Men and Women's Apparel	Craig French	Hollywood, CA	http://crookedjawfashions.com/	200000.0	20.0	1000000.0	1.0	Barb Corco
7	False	Attach Noted is a detachable "arm" that holds ...	2.0	Productivity Tools	Mary Ellen Simonson	Gardena, CA	http://www.attachnoted.com/	100000.0	20.0	500000.0	1.0	Barb Corco
8	False	A safety device for seatbelts. It prevents the...	2.0	Automotive	Robert Alison	Las Vegas, NV	http://www.nobucklenostart.com/	500000.0	10.0	5000000.0	1.0	Barb Corco

Figure 33: Not secured data

3. **The following exercise is to be done for both the corpora:**

   a) **Find the number of characters for both the corpuses.**
   The total number of characters for both the corpuses is calculated as follows:-

```
Total number of characters in Not Secured Corpus: 34899
Total number of characters in Secured Corpus: 50302
```

   b) **Remove Stop Words from the corpora. (Words like 'also', 'made', 'makes', 'like', 'this','even' and 'company' are to be removed)**

   After removing stop words the new corpora will look like

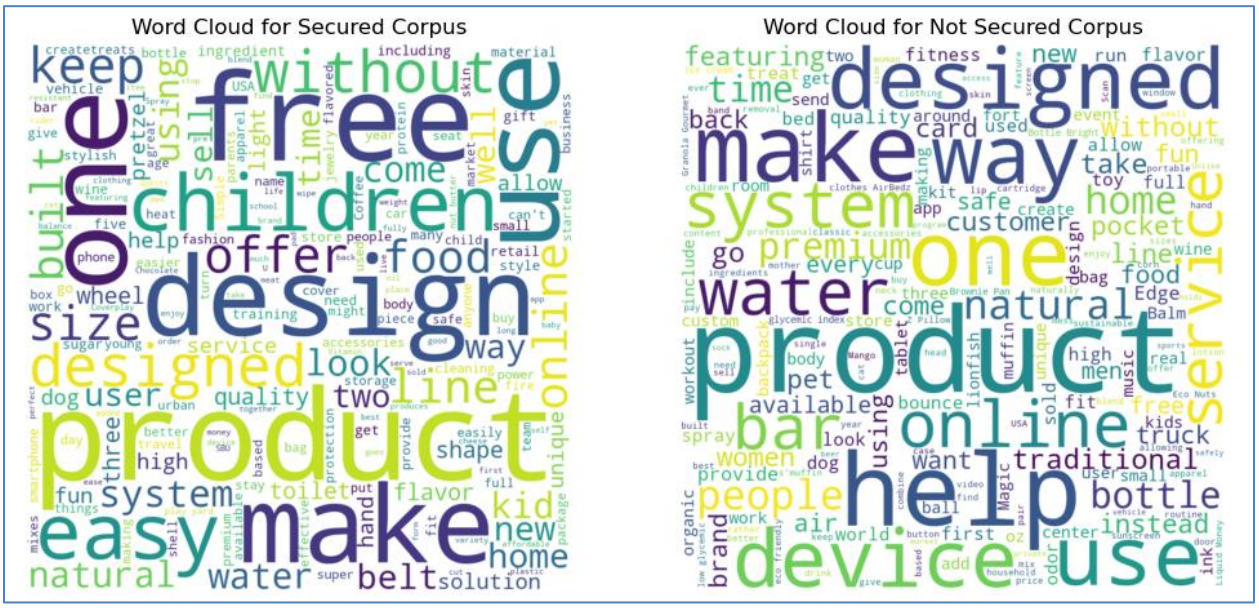
```
Secured Corpus (After removing stop words):
1       Retail wholesale pie factory two retail locati...
2       Ava Elephant godsend frazzled parents young ch...
5       One first entrepreneurs pitch Shark Tank, Susa...
12      line books written help children find inner calm.
16      Coverplay slipcover children's play yards. Muc...
                              ...
489     SynDaver Labs synthetic body parts use medical...
490     Zoom Interiors virtual service interior design...
491     Spikeball started casual outdoors game, grown ...
492     Shark Wheel literally reinvent wheel. innovati...
494     Sway Motorsports three-wheeled, all-electric, ...
Name: description, Length: 204, dtype: object

Not Secured Corpus (After removing stop words):
3       Organizing, packing, moving services delivered...
4       Interactive media centers healthcare waiting r...
6       mixed martial arts clothing line looking becom...
7       Attach Noted detachable "arm" holds Post-It no...
8       safety device seatbelts. prevents driver start...
                              ...
482           Buck Mason high-quality men's clothing USA.
484     Frameri answers question, "Why glasses flexibl...
485     Paleo Diet Bar nutrition bar gluten, soy, dair...
488     Sunscreen Mist adds another point access sunsc...
493     Adriana Montano wants open first Cat Cafe Flor...
Name: description, Length: 183, dtype: object
```

   c) **What were the top 3 most frequently occurring words in both corpuses (after removing stop words)?**

```
Top 3 most frequent words in secured corpus (after removing stop words): [('easy', 16), ('make', 16), ('without', 15)]
Top 3 most frequent words in not secured corpus (after removing stop words): [('make', 15), ('system', 11), ('products', 10)]
```

d) Plot the Word Cloud for both the corpora.

The word cloud for both the copra is shown below:-



4. Refer to both the word clouds. What do you infer?

The word 'device','Service','Home' etc. appears prominently in the word cloud associated with deals not secured, indicating that it occurred frequently in discussions or documents related to rejected deals. Conversely, the word 'device','Service','Home' is not easily found in the word cloud associated with deals secured. This suggests that the presence of these words may be associated with unsuccessful deal outcomes.

5. Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?

Based on the analysis it is true and it appears that entrepreneurs who introduced devices may indeed be less likely to secure a deal. The frequency of the word 'device' in the word cloud associated with deals not secured implies that its presence may be a contributing factor to deal rejection. However, it's important to note that correlation does not imply causation, and there could be various other factors influencing deal outcomes.